STAC67: Regression Analysis

Assignment 1
(Total: 100 points)

Please submit R Markdown file for Q. 1 and Q. 6 along with your submission of the assignment.

Q. 1 (10 pts) This question is to practice R to sample from a Normal distribution. Obtain random samples from a Normal with mean $\mu = 100$, $\sigma = 20$ of size n = 100, 1000, 10,000, 100,000.

When you generate a random number, use R code, **set.seed(your student number)** before the R codes of generating a random number, so that we can replicate the result.

(a) (5 pts) On a single page (2 rows, 2 columns) give the histograms on the same set of bins, with a normal density superimposed on each. Comment on the approximation accuracy.

(b) (5 pts) For each sample size, give the mean, standard deviation, and the following percentiles (2.5, 25, 50, 75, 97.5). Compare these with the theoretical values.

Q. 2 (14 pts) (a) (4 pts) Prove the following equalities.

(i) $S_{XX} = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$

(ii) $S_{XY} = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$

(b) Suppose that $(Y_1, X_1), \ldots, (Y_n, X_n)$ is a data set to which we fit a simple linear regression. Let $\hat{\beta}_1$ be the least squares estimate of the slope with Y and let r be the sample correlation coefficient.

(i) (5 pts) Show that

$$\hat{\beta}_1 = r\frac{s_Y}{s_X}$$

where s_Y and s_X are the sample standard deviations of $Y_1 \ldots Y_n$ and $X_1 \ldots X_n$ respectively.

(ii) (5 pts) Show that

$$\frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Q. 3 (16 pts) (4 pts each) Anastrozole is a drug often used to treat breast cancer patients. One study attempted to see if the effect of Anastrozole is associated with the age of patients. The response variable Y is the change the levels of cortisol-binding globulin (CBG). and the covariate x is age. The following summary statistics were reported.

$$\begin{array}{lll} n = 26 & \sum X_i = 1613 & \sum Y_i = 281.9 \\ S_{XX} = 3756.96 & S_{YY} = 465.34 & S_{XY} = -757.64 \end{array}$$

(a) Find the least squares estimates of the intercept and slope.

(b) Give the standard errors for your estimates in (a).

(c) Construct 95% confidence intervals for the true intercept and true slope.

(d) What conclusions would you draw from your results?

Q. 4 (28 pts) (4 pts each) We fit the linear regression model without the intercept, $Y_i = \beta_1 X_i + \epsilon_i, i = 1, \dots n$,

(a) Find the least square estimator of β_1.

(b) Denote the estimator by $\hat{\beta}_1$ then the estimated model is $\hat{Y}_i = \hat{\beta}_1 X_i$. Let $e_i = Y_i - \hat{Y}_i$. Can you conclude $\sum_{i=1}^{n} e_i = 0$?

(c) Assume that the error term are independent and identically distributed, $N(0, \sigma^2)$ with σ^2 unknown for successive questions (c to f). Find the Standard Error for the estimator of β_1.

(d) Design a procedure to test

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

(e) (5 pts) The data is collected for six observations.

i :	1	2	3	4	5	6
X_i :	7	12	4	14	25	30
Y_i :	128	213	75	250	446	540

Find the maximum likelihood estimator of β_1 and evaluate its value.

(f) Consider another estimator $\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i}$. Derive its mean and variance.

(g) Which estimator has the smaller variance? Why?

Q. 5 (8 pts) Consider a simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \text{ for i} = 1, 2, \cdots, \text{n},$$

where $\epsilon_i \sim N(0, \sigma^2)$ and

$$\hat{\beta}_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$$

is the least squares estimator of β_1. Imagine $\hat{\beta}_1^* = \sum_{i=1}^{n} c_i Y_i$ is any other unbiased estimator of β_1 with c_i being arbitrary constant. Prove that $Var(\hat{\beta}_1) \leq Var(\hat{\beta}_1^*)$. That is, prove that the least squares estimator of β_1 has the minimum variance among all other linear unbiased estimators of β_1.

Q. 6 (24 pts) (4 pts each) For this question, use R Markdown file. The data set, "vote.txt" is posted at Quercus. The data contains the incumbent party's vote percentage of the two-party vote coded as **vote** and average personal income growth in the previous years coded as **growth**. The political scientist Douglas Hibbs forecasts elections based solely on economical growth.

(a) Obtain a scatter plot between two variables (make sure which variable goes to y axis), also add the fitted linear regression line.

(b) Fit a simple linear regression in R, predicting elections from the economy. Interpret both estimates ($\hat{\beta}_0$ and $\hat{\beta}_1$) in words.

(c) Predict the incumbent party's vote in 2008 election and based on that, who will won the election between "McCain" and "Obama"? (both by hands and in R)

(d) Test whether there is a positive association between incumbent party's vote share and economical growth.

(e) Give a 95% confidence interval for the mean incumbent party's vote share change as economical growth increases in one unit (percent) (both by hands and in R).

(f) Compute the probability that $P(|\hat{\beta}_1 - \beta| > 1)$.